数据数据-Task2

2) 简略观察数据(head()+shape)

Train_data.head().append(Train_data.tail())

	SaleID	name	regDate	model	brand	bodyType	fuelType	gearbox	power	kilometer	...	v_5	v_6
0	0	736	20040402	30.0	6	1.0	0.0	0.0	60	12.5	...	0.235676	0.10
1	1	2262	20030301	40.0	1	2.0	0.0	0.0	0	15.0	...	0.264777	0.12
2	2	14874	20040403	115.0	15	1.0	0.0	0.0	163	12.5	...	0.251410	0.11
3	3	71865	19960908	109.0	10	0.0	0.0	1.0	193	15.0	...	0.274293	0.11
4	4	111080	20120103	110.0	5	1.0	0.0	0.0	68	5.0	...	0.228036	0.07
149995	149995	163978	20000607	121.0	10	4.0	0.0	1.0	163	15.0	...	0.280264	0.00
149996	149996	184535	20091102	116.0	11	0.0	0.0	0.0	125	10.0	...	0.253217	0.00
149997	149997	147587	20101003	60.0	11	1.0	1.0	0.0	90	6.0	...	0.233353	0.00
149998	149998	45907	20060312	34.0	10	3.0	1.0	0.0	156	15.0	...	0.256369	0.00
149999	149999	177672	19990204	19.0	28	6.0	0.0	1.0	193	12.5	...	0.284475	0.00

哈哈哈哈，这个还真的是第一次见，原来还可以头+尾一起展示

通过以上两句可以很直观的了解哪些列存在 “nan”，并可以把nan的个数打印，主要的目的在于 nan存在的个数是否真的很大，**<u>如果很小一般选择填充，如果使用lgb等树模型可以直接空缺，让树自己去优化，但如果nan存在的过多、可以考虑删掉</u>**

可以发现除了notRepairedDamage 为object类型其他都为数字 这里我们把他的几个不同的值都进行显示就知道了

```
Train_data['notRepairedDamage'].value_counts()
```

```
[21]: 0.0    111361
      ,-       24324
      ,1.0     14315
      ,Name: notRepairedDamage, dtype: int64
```

可以看出来‘ - ’也为空缺值，因为很多模型对nan有直接的处理，这里我们先不做处理，先替换成nan

```
Train_data['notRepairedDamage'].replace('-', np.nan, inplace=True)
```

```
Train_data['notRepairedDamage'].value_counts()
```

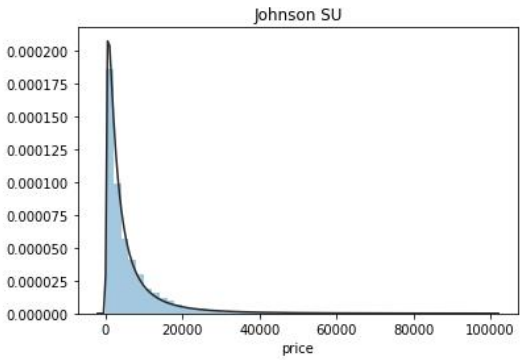
```
[23]: 0.0    111361
      ,1.0     14315
      ,Name: notRepairedDamage, dtype: int64
```

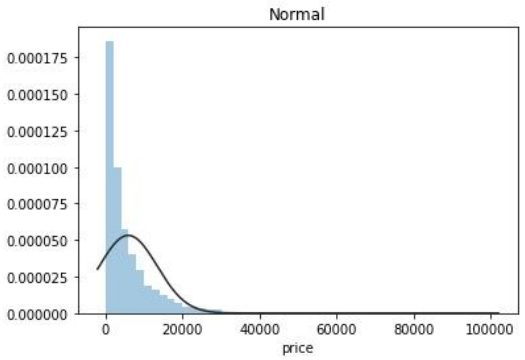
```
Train_data.isnull().sum()
```

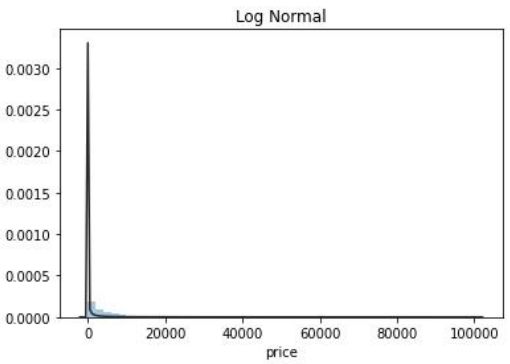
感觉这里对于nan的处理比较好

```
## 1) 总体分布概况（无界约翰逊分布等）
import scipy.stats as st
y = Train_data['price']
plt.figure(1); plt.title('Johnson SU')
sns.distplot(y, kde=False, fit=st.johnsonsu)
plt.figure(2); plt.title('Normal')
sns.distplot(y, kde=False, fit=st.norm)
plt.figure(3); plt.title('Log Normal')
sns.distplot(y, kde=False, fit=st.lognorm)
```

```
[32]: <AxesSubplot:title={'center':'Log Normal'}, xlabel='price'>
```







价格不服从正态分布，所以在进行回归之前，它必须进行转换。虽然对数变换做得很好，但最佳拟合是无界约翰逊分布

```
## 2) 查看skewness and kurtosis
sns.distplot(Train_data['price']);
print("Skewness: %f" % Train_data['price'].skew())
print("Kurtosis: %f" % Train_data['price'].kurt())
```

Skewness: 3.346487

第一次听说这个无界约翰逊分布

一.偏度（Skewness）

Definition:是描述数据分布形态的统计量，其描述的是某总体取值分布的**对称性**，简单来说就是数据的不对称程度。。

偏度是三阶中心距计算出来的。

（1）Skewness = 0 ，分布形态与正态分布偏度相同。

（2）Skewness > 0 ，正偏差数值较大，为正偏或右偏。长尾巴拖在右边，数据右端有较多的极端值。

（3）Skewness < 0 ，负偏差数值较大，为负偏或左偏。长尾巴拖在左边，数据左端有较多的极端值。

（4）数值的绝对值越大，表明数据分布越不对称，偏斜程度大。

计算公式：

Skewness=E[((x-E(x))/(\sqrt{D(x)}))^3]

| Skewness| 越大，分布形态偏移程度越大。

以前做kaggle用到过，一般偏度0.99的feature我就直接去掉了

二.峰度（Kurtosis）

Definition:偏度是描述某变量所有取值分布形态陡缓程度的统计量，简单来说就是数据分布顶的**尖锐程度**。

峰度是四阶标准矩计算出来的。

（1）Kurtosis=0 与正态分布的陡缓程度相同。

（2）Kurtosis>0 比正态分布的高峰更加陡峭——尖顶峰

（3）Kurtosis<0 比正态分布的高峰来得平台——平顶峰

计算公式：

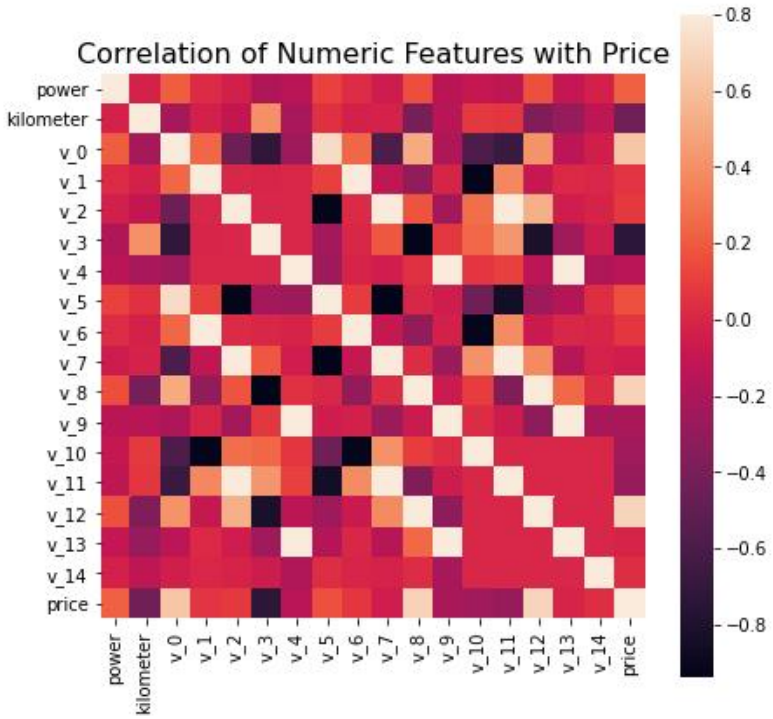
Kurtosis=E[((x-E(x))/ (\sqrt(D(x))))^4]-3

```
f , ax = plt.subplots(figsize = (7, 7))

plt.title('Correlation of Numeric Features with Price',y=1,size=16)

sns.heatmap(correlation,square = True,  vmax=0.8)
```

```
[48]: <AxesSubplot:title={'center':'Correlation of Numeric Features with Price'}>
```



heatmap，我一般会选择相似度>99%的两个feature中去掉一个

可视化是我的短板，一定要好好学学

https://www.jianshu.com/p/6e18d21a4cad

2.4 经验总结

所给出的EDA步骤为广为普遍的步骤，在实际的不管是工程还是比赛过程中，这只是最开始的一步，也是最基本的一步。

接下来一般要结合模型的效果以及特征工程等来分析数据的实际建模情况，根据自己的一些理解，查阅文献，对实际问题做出判断和深入的理解。
最后不断进行EDA与数据处理和挖掘，来到达更好的数据结构和分布以及较为强势相关的特征

数据探索在机器学习中我们一般称为EDA（Exploratory Data Analysis）：

> 是指对已有的数据（特别是调查或观察得来的原始数据）在尽量少的先验假定下进行探索，通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。

数据探索有利于我们发现数据的一些特性，数据之间的关联性，对于后续的特征构建是很有帮助的。

1. 对于数据的初步分析（直接查看数据，或.sum(), .mean()，.descirbe()等统计函数）可以从：样本数量，训练集数量，是否有时间特征，是否是时许问题，特征所表示的含义（非匿名特征），特征类型（字符类似，int，float，time），特征的缺失情况（注意缺失的在数据中的表现形式，有些是空的有些是”NAN”符号等），特征的均值方差情况。
2. 分析记录某些特征值缺失占比30%以上样本的缺失处理，有助于后续的模型验证和调节，分析特征应该是填充（填充方式是什么，均值填充，0填充，众数填充等），还是舍去，还是先做样本分类用不同的特征模型去预测。
3. 对于异常值做专门的分析，分析特征异常的label是否为异常值（或者偏离均值较远或者事特殊符号）,异常值是否应该剔除，还是用正常值填充，是记录异常，还是机器本身异常等。
4. 对于Label做专门的分析，分析标签的分布情况等。
5. 进步分析可以通过对特征作图，特征和label联合做图（统计图，离散图），直观了解特征的分布情况，通过这一步也可以发现数据之中的一些异常值等，通过箱型图分析一些特征值的偏离情况，对于特征和特征联合作图，对于特征和label联合作图，分析其中的一些关联性。

收获很大，好多可视化图是我第一次接触，值得反复阅读